Design and Implementation Assessment Device (Version 0.3)

High quality research in education is in demand. Rising standards, accountability requirements, and legislation (e.g., No Child Left Behind) all suggest that education policies and practices should be based on sound scientific evidence. Standards of evidence for judging the effectiveness of educational interventions and approaches should be clearly defined, broadly accepted, and used by researchers, policy makers, and practitioners.

The What Works Clearinghouse (WWC) was established to help improve the education of American students by

- producing high quality summaries of research on the effects of educational interventions and approaches on student outcomes
- promoting the use of rigorous scientific methods in studies of educational effectiveness
- promoting the use of rigorous research in education decision-making, and
- facilitating public and educator access to research-related resources.

The WWC is committed to approaching this work in a systematic fashion that is open to critical appraisal and suggestions for improvement.

As part of its mission, the WWC will produce syntheses of research summarizing the evidence pertaining to the effectiveness of educational interventions and approaches. When enough high-quality evidence exists, the syntheses will (a) estimate the magnitude of the effect of the intervention or approach on achievement-related variables and (b) attempt to identify variations in interventions and approaches that are associated with success. To this end, the WWC is in the process of establishing a way to assess the design and implementation of research that attempts to draw causal inferences about the effectiveness of educational interventions and approaches.

Assessment of study design and implementation

Assessments of the design and implementation of research studies can be used in multiple ways. The first way is to establish criteria for including and excluding studies from a research synthesis. Specifically, there may be some studies that are so poorly designed and implemented that we would not want to include them in our evidence reports. Their results are so suspect that we would not know what to conclude from them. A second way we use design and implementation assessments is to examine the impact of different research designs on conclusions and see if different designs and implementations lead to different results. Finally, design and implementation assessments can be used to help draw conclusions about the cumulative strength of an entire set of studies. Each different way of using assessments will play a role in how the What Works Clearinghouse will use its research design and implementation assessments.

A brief review of previous attempts to assess study design and implementation. Before setting out to develop our own device to assessment study design and implementation, we examined devices that had been developed by others. We found several problems with the devices that we reviewed. One problem was that they frequently differ in terms of the number of features of a study that they consider important. And, when they do have the same number and include the same features, they vary greatly in the amount of importance that they place on each feature. For example, one scale reviewed by Jüni

et al. (1999) weighted the use of random assignment of participants to conditions as worth 15% of the total score while in another scale this feature was only weighted 4%. The same two scales nearly reversed the weight that they placed on whether the participants knew which condition of the experiment they were in. Obviously, when weights vary it is plausible that the same study will get different scores on the different assessment devices.

In addition, when a single score for studies is generated by a scale, it becomes very difficult to figure out how much importance is being given to each of the different study features. For example, hypothetical Scale A might give Study X a low internal validity score (say "20") but a high external validity score (say, "60") to get a total score of "80." Hypothetical Scale B might give the same Study X a high internal validity score ("60") but low external validity ("20") that would result in the same total score of "80." When multiple dimensions are combined it makes it very difficult to understand what those scores mean. So, the WWC design and implementation scale does not come up with a single number but keeps independent dimensions of design and implementation separate.

Assumptions guiding the development of an assessment of study design and implementation

In order to give you an idea about some of the issues we considered before we began developing the WWC's Design and Implementation Assessment Device (DIAD) , we tried to make explicit the assumptions that we used in constructing it. Four assumptions underlie the development of the WWC DIAD.

The first assumption is that *design and implementation varies across studies in ways that affect the confidence that can be placed in their results*. This is simply a way of saying that studies vary in the confidence that we have in their results and that this is something that can be measured. Obviously, if we do not make this assumption, there is no reason to develop a scale.

Our second assumption is that *a study's design and implementation has more than one dimension*. The WWC DIAD does not come up with a single number to represent each study. Instead, we have four judgments at the most general level and we rely on a minimum of eight different judgments at the operational level.

A third assumption is that *studies with different purposes require different design and implementation assessments*. For example, some studies attempt to determine the effectiveness of different educational interventions and approaches. Other studies involve assessing the validity of tests and measurements. Clearly, deciding whether or not a study has done a good job of assessing the validity of a test or measurement will require looking at many different criteria and issues than will a study of an intervention's effectiveness. A third type of studies can focus on qualitative descriptions, or process and implementation issues, are going to invoke issues of design and implementation which are different than the other purposes.

We do not believe that the instrument we describe below is appropriate for all types of research with all types of purposes. This instrument focuses on effectiveness research, that is, research that examines the effects and effectiveness of an intervention or an approach Other research purposes would require different kinds of assessment devices.

Finally, our fourth assumption is that *Donald Campbell's approach to assessing the validity of inferences that arise from a study's design and implementation is a good way to think about the*

certainty we can have in a study's conclusions. We are not suggesting that this is the *only* framework for understanding research design and implementation, but, for us, it is an obvious and natural choice.

Principles guiding the development of an assessment of study design and implementation

Next, we would like to make explicit the guiding principles used in developing our design and implementation device. As you examine this device and suggest modifications, it is important to understand what our guiding principles were. You will see that sometimes the guiding principles might lead to different choices about how to develop the DIAD. Quite often -- more often than we would have hoped! -- we had to make a decision about which one of our guiding principles was paramount at a particular decision point.

Our first principle was that *a design and implementation scale should be understandable by a broad audience, including policy makers and practitioners*. This principle obviously is very critical to the What Works Clearinghouse effort. However, the principle creates a very difficult and challenging task for us, albeit one that is very exciting. It forces us to think about how to express what research quality means to a broad audience, but, at the same time, to do so in a manner that meets the standards of systematicity, rigor, and transparency that methodologists require. You will see that this is the first and most important principle that guided many of decisions we made.

Our second guiding principle was that *agreement on a design and implementation scale's contents should approach consensus*. We hoped to come up with a set of scales or subscales that most people would agree properly define and measure issues related to study design and implementation. However, we use the term "approach consensus" because we know that we will not attain complete consensus. Doing research in applied settings is very complex and people come to it with different beliefs about what aspects of design are most important. There will always be some disagreement. However, we do believe that consensus is important and we strove wherever possible to develop an instrument that would engender the greatest degree of agreement across the broadest community of educational researchers.

Our third assumption was that *assessments of design and implementation should be tied explicitly to the way the studies were carried out*. This is one of the characteristics that sets our device apart from many of the other devices that we examined. As we discuss the specifics of the WWC DIAD, you will see that we attempted, whenever possible, to ground our assessments of study design and implementation on concrete, low inference judgments about how studies were conducted. Many scales of study quality make broad assessments of abstract characteristics of studies. We tried to make those assessments as explicit as possible, and then build more abstract judgments based upon those concrete assessments.

Our fourth guiding principle was that *design and implementation scales should be based on criteria that are defined transparently*. This is also a goal that is more difficult to realize than it is to set out. However, we attempted to make certain that whoever applies the scale will do so in a manner in which their judgments are open to inspection by others, so we can disagree, critique, and argue in a precise and constructive manner.

Our fifth guiding principle was that *a design and implementation scale should include commonly held ideas about what constitutes good and poor research practice*. Some of these judgments are arrived at more easily than others, but it is important, if we want educational practices to be based on

good research, to make explicit and employ consensual understandings about what it is good for studies to do and what is bad for studies to do. Many times in the course of drafting the scale, we referred to textbooks, both advanced and introductory, regarding how to conduct social research. The sense of these books is that there are better and worse ways to design and implement studies of the outcomes of interventions and approaches. We think that commonly shared scientific principles exist and it is important that they be embedded in our design and implementation scale.

Our sixth guiding principle was that *a design and implementation scale should permit some flexibility for adaptation based on the characteristics of a topic area*. Complementing the importance of the shared scientific ideals, it is also the case that each study comes embedded within its own context. Thus, we have to balance the notion of the shared scientific ideals, which transcend all research, with realizations of how those ideals will be carried out in any particular study. Each study will have uniqueness. There has to be flexibility in the system that permits the people who are applying the device to be able to reflect in it what makes for a good study within their particular context. Thus, there are places in the device where the team leaders preparing a WWC Evidence Report are asked to provide contextual information that feeds into and assists in its completion.

Our seventh guiding principle was that *a design and implementation scale should be uniform in format and presentation*. We attempted to create a system of responses that could be applied across a wide range of educational effectiveness studies. The design and implementation scale has standardized output.

Our eighth and final guiding principle was that *studies that fail to report important design features have greater potential for poor quality*. Often, one of the major problems in assessing research, whether formally or informally, is that we simply don't know what happened. We think it's very important for users of research to get detailed research reports. When information is reported, we have built into our device a way for us to be able to code that information. But when the information is unreported, we typically hold that against the study, unless it makes little sense to do so.

The Structure of the DIAD

With these assumptions and guiding principles in mind, we developed the What Works Clearinghouse Design and Implementation Assessment Device, or DIAD. We attempted to create one instrument that can be used to answer questions at three different levels of generality. We wanted the most general level to be understandable to an audience of non-researchers and the most detailed level to be specific enough to satisfy researchers' desire for comprehensiveness and explicitness.

The DIAD is largely written in the language of two group comparisons. We expect that our initial reviews will be on topics in which this type of design predominates. However, other types of designs (e.g., regression discontinuity, multiple baseline time series) could be incorporated into the structure of the DIAD with minimal adjustment. Other designs and analyses meant to uncover causal relations (e.g., instrumental variable analysis, single-subject designs) will likely require attention to somewhat different sets of issues. The WWC will make modifications to the DIAD to accommodate these designs as needed.

The device is arranged as a hierarchy of related questions (and answers) starting with specific study characteristics that are used to build global assessments of the *certainty* we have that a study has uncovered the effects of the intervention or approach. The questions are arranged so that answers at

one level—the most specific level—feed into a set of composite questions at a second level, and then into a third level of even more general questions. In essence, anybody could look at the device at three different levels of abstractness depending upon the uses they intended for its assessments. But, no matter which level served their purpose, the user could also see the other levels and know how the assessments at each level related to or were dependent upon the level below it.

At the most specific level of the DIAD there are questions about specific design and implementation characteristics of the study. There are approximately 50 questions that are answered about each study, and each question is answered in a "yes" or "no" format. For example, at this level you will find a question that reads "Were participants randomly assigned to conditions?" The study coder would use and either “yes” or “no” to answer this question. (If the authors of the study did not report how participants were allocated to groups, that study would be given a "no" on the random assignment question). Another question that appears at the most specific level is "Were comparison groups or participants aware that they were in the comparison group?" This question relates to issues concerning how comparison groups might have reacted to not receiving the intervention. Again, the question is answered either "yes" or "no."

The answers to specific questions are then compared to a set of algorithms used to generate answers to eight composite questions about design and implementation. These eight questions are more general. For example, one question at this level is "Were the participants in the policy or practice group comparable to the participants in the comparison group"? The answer to the more specific question mentioned above on random assignment clearly “feeds into” answering this question, but other questions are relevant as well.

After the eight composite questions are answered, they can be used to answer four more global questions about a study. Again, explicit algorithms are used to turn the composite eight questions into the global four questions. The questions at the most global level deal with what Cook & Campbell (1979), and Shadish, Cook & Campbell, (2002) referred to as the four most general sets of threats to the validity of a study: construct validity, internal validity, external validity, and statistical validity. The four questions that the DIAD answers at its most global level are related to each of these four sets of threats. In order, they are

1. “Were the intervention or approach, and outcome, properly defined?”
2. “Was the intervention or approach the cause of the change in the outcome?”
3. “Was the intervention or approach tested on relevant participants (for example, students and schools), and environments (for example, classrooms and occasions)?”
4. “Could accurate effect sizes be derived from the study report?”

At both the composite question level and the global question level, there are four possible answers to each question, “yes”, “maybe yes”, “maybe no” and “no”. Which answer a specific study gets to each of the composite and global questions depends on the answers that were given to the questions “below” that relate to it.

The eight composite questions are embedded within the four global questions in that the responses to each of the composite questions contribute to the responses to the global questions. For example, the answer to the global question about *internal validity* (“Was the policy or practice the cause of the change in the outcome”?) has associated with it the answer to the composite question on whether participants in the policy or practice group were comparable to the participants in the comparison

group. But, another composite question feeds into this answer as well. This question relates to treatment contamination, or confusion between what specifically happened to the intervention and the comparison group.

Next, we will demonstrate how hierarchical arrangement will work for one part of one of the DIAD, specifically, the composite level question that deals with selection bias. The composite question that relates to selection bias is "Were the participants (students or schools) in the policy or practice group comparable to the participants in the comparison group?" The DIAD has four answers to this question:

1. "Yes, participants were randomly assigned to conditions and there was no differential attrition or severe overall attrition"
2. "Maybe yes, EITHER random assignment was used but there was differential and/or severe overall attrition OR although random assignment was not used, there was no attrition problem and reasonable steps were taken to make the groups comparable"
3. "Maybe no, random assignment was not used and although steps were taken to make the groups comparable, they did not appear to be adequate", and
4. "No, it is highly unlikely that the participants in the groups were comparable."

Each answer is associated with a particular pattern or patterns of responses to the more specific questions. There are six specific questions that relate to selection bias. The first two deal with how the participants were allocated to groups: (a) "Was there random assignment?" and (b) "Was there differential or severe attrition?" Next, there are four questions that deal with equating the intervention and comparison groups: (a) "Were post-intervention procedures used that equated groups on a pre-test of the outcome?" (b) "Were pre-intervention procedures used that equated groups on a pre-test of the outcome?", (c) "Were post-intervention procedures used that equated groups on other important variables?", and (d) "Were pre-intervention procedures used that equated groups on other important variables?"

Finally, we need to point out two issues related to how different design features are defined. First, many terms we use are defined in a glossary that accompanies the DIAD. Again, these are draft definitions and some definitions may change. New terms may be added. In the DIAD you can identify these terms because they will be underlined. Please, let us know if you have suggestions regarding these definitions. Second, we noted above that a good design and implementation assessment device needs to be flexible. In particular, there will be certain terms and contextual issues that are specific to each topic area. Therefore, you will see in the DIAD some terms that appear quite vague, or ambiguous. In most instances these are terms that will be given more specific meaning by the leaders of the teams doing the WWC Evidence Reports. Specifically, prior to beginning an Evidence Report and completing the DIAD for any study, the leaders of the Evidence Report team will answer a list of questions pertinent to the Evidence Report. For each question of this type, we have italicized and starred (*) the term that appears in the DIAD.

1. For purposes of sampling, what constitutes the *local pool of participants**?
2. For research on this topic how would you define *differential attrition from the intervention and control groups**?
3. For research on this topic, how would you define *severe attrition overall**?
4. What constitutes a *minimal sample size** that would permit a sufficiently precise estimate of the effect size?

5. What are the *important characteristics* that define the *target population**?
6. What are the *important characteristics** of participants that might be related to the policy or practice effect that ought to be equated if a study does not employ random assignment?
7. What *characteristics of subgroups of participants and settings** are important to test within a study to determine whether a policy or practice is effective within these groups?
8. What *commonly-shared and/or theoretically derived ideas about the policy or practice** should be reflected in its definition and implementation?
9. What are the *important classes of outcomes**?
10. What aspects of the outcomes do *commonly-shared and/or theoretically derived ideas** suggest should be part of their definition?
11. Was the study conducted during the *time frame appropriate for the Evidence Report**?
12. What is the *appropriate interval for measuring the treatment effect** relative to the end of the intervention?
13. What *statistical properties of the data** are important to obtain an accurate estimate of an effect size?

While conducting a WWC Evidence Report, the investigators will largely need to rely on the labels provided by the authors of the primary studies. For example, suppose a study states that the sample was comprised of "at-risk males aged 11-13" but provides no information about how the determination of "at-risk" was made. In cases such as this, the label provided by the study author is assumed to be correct. If a report does provide information about how labels were arrived at, that information will be compared to the Evidence Report team's definition of the category. Similarly, in most cases investigators will have to assume that the studies they review are free of mistakes that are not observable at the level of the report (e.g., if the codes assigned to represent participant sex were accidentally reversed). Of course, if information in the report does suggest a mistake was made, this will be noted.

Another issue concerns how the DIAD will be completed when a study includes multiple interventions or approaches, control groups, and/or outcome measures. With regard to multiple groups, we anticipate that most of the design and implementation assessments will remain consistent across all groups (e.g., the use of random assignment) but when they do differ (e.g., multiple interventions that are described with more or less precision) we would expect that the evidence report team will provide separate judgments for each intervention or approach (e.g., "for Intervention X this study provided clear definitions but for Intervention Y this study was less clear"). With regard to multiple outcome measures, we expect that when a study includes both relevant and irrelevant measures, the evidence report team will exclude the irrelevant measures. When studies have multiple relevant measures that vary in precision, we expect that the report team will judge the study on the basis of its best measures but test whether the precision of the measure was associated with study outcomes as part of the research synthesis.

The questions on the DIAD do not represent the totality of information that will be extracted from studies included in WWC Evidence Reports. The review teams will also extract more specific information about (a) the intervention or approach and how it was implemented, (b) the research design and how it was implemented, (c) the individuals in the study, and (d) the outcomes of the study.

Finally, in addition to the questions on the DIAD, the team leaders will be provided with a list of other desirable characteristics of study design and implementation that are not part of the DIAD

(Evidence Report team leaders may also add to this list). These may be desirable but not dispositive aspects of methodology or might be of particular relevance to particular research questions. As an example, even when random assignment is not possible, it is often desirable to allocate participants to study groups on a basis that is *not* related to the outcome variable. These characteristics of studies will be coded and, if possible, examined as potential moderators of effect size.

We hope you found this introduction to the DIAD helpful. We refer to this version as DIAD Version 0.3. We call it 0.3 because it is a draft. We need input from others to insure that the assessment device used by the What Works Clearinghouse best meets our goals and principles. Your comments will be valuable. Please contact us at wwcinfo@w-w-c.org. Please include the phrase "Comments on DIAD" in the subject line.

How to Read DIAD (Version 0.3)

The following pages detail the initial, draft questions that comprise the Design and Implementation Assessment Device (DIAD). The DIAD will be used primarily as a screening device to (a) ensure that studies included in WWC Evidence Reports meet explict scientific standards and (b) assist in assessing whether the results of studies included in WWC Evidence Reports are related to study design and implementation. In addition, we hope that the DIAD will prompt more critical assessments of educational research and an improvement in educational research practice.

As described above, the DIAD results in three levels of assessment of design and implementation that are hierarchically related. The persons completing the DIAD answer highly specific "yes" or "no" questions. These are then compared to algorithms that result in the answers to eight more general composite questions about design and implementation. Finally, the answers to the eight questions combine yet again to answer four even broader questions.

At the most global level, the DIAD poses four questions that should be understandable (with some minimal additional explanation) to educational policy makers and practitioners. The four questions at this level reflect the four major classes of Campbell's "threats to validity" framework (Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). These questions are:

- Were the *intervention or approach** and outcome properly defined?
- Was the *intervention or approach** the cause of the change in the outcome?
- Was the *intervention or approach** tested on targeted participants (for example, students, schools) and environments (for example, classrooms, occasions)?
- Could accurate effect sizes be derived from the study report?

The next level of the DIAD poses eight questions and is aimed at an audience with a somewhat higher level of sophistication with regard to research methodology and statistics. These eight questions are presented below, along with four possible answers to each question. Each answer starts with "yes", "maybe yes", "maybe no", or "no". Then, we present tables that include the highly specific "yes" or "no" questions associated with each of the eight composite questions. The tables also contain the algorithms that are applied to the specific questions in order to generate the answers to the eight questions. Finally, for each of the eight questions we propose which answers will result in the inclusion of a study in WWC Evidence Reviews.

Three clarifying points about the tables are necessary. The first relates to how the charts should be read. To begin with, some cells are irrelevant given the patterns of answers to other questions. As an example, part of one table looks like this:

Construct Validity – Intervention or Approach	**Yes**
1.1a Was the *intervention or approach** described at a level which would allow its implementation by other implementers?	Yes
1.1b If no, was the *intervention or approach** a member of a broader class (across which significant variation in content can be expected)?	n/a

Question 1.1b has an "n/a" in the cell because the "yes" answer to question 1.1a makes this question irrelevant. In other cases, either a "yes" or a "no" answer to a question would result in the same overall

answer (such as when the question isn't terribly important to the overall conclusion). In these cases, we've entered "Yes or No" in the appropriate cell.

The table below presents 6 questions that assess selection bias. The composite question that these questions feed into is: "Were the participants (e.g., students, schools) in the group receiving the intervention or approach comparable to the participants in the comparison group?"

Internal Validity – Selection	**Yes**	**Maybe yes**	**Maybe yes**	**Maybe yes**	**No**	**If unreported:**
3.1 Was random assignment used to assign participants to intervention/approach and comparison groups?	Yes	Yes	No	No	No	No
3.2 Was there differential attrition* between intervention/approach and comparison groups or severe attrition* overall?	No	Yes	No	Yes	Yes or No	No
3.3 If the groups were not demonstrably equivalent, were post-intervention procedures used that equated groups on a pretest of the outcome?	Yes or No	Yes or No	Yes or No	Yes	No	No
3.4 Were pre-intervention procedures used that equated groups on a pretest of the outcome?	Yes or No	Yes or No	Yes	Yes or No	No	No
3.5 Were post-intervention procedures used that equated groups on other important variables?	Yes or No	Yes or No	Yes or No	Yes	No	No
3.6 Were pre-intervention procedures used that equated groups on other important variables?	Yes or No	Yes or No	Yes	Yes or No	No	No

Note: A pattern of answers to these questions that is not specifically identified results in a "Maybe No."

Each column above represents a pattern of answers to the specific questions that will result in an answer to the composite question on selection bias. For example, if a study used random assignment and had no differential attrition and no severe attrition, the study gets a "yes" answer to the question "Were the participants (e.g., students, schools) in the group receiving the intervention or approach comparable to the participants in the comparison group?" There are three ways for the question to be answered "maybe yes"; one for a study that used random assignment but experienced differential or severe attrition, and two for studies in which there was no random assignment.

Second, the last column of all but one table is labeled "If Unreported." The "If Unreported" column lists proposed default options for coding information that is not readily discernable from the report or related, easily obtainable documents. As an example, if the mechanism used to assign participants to groups could not be determined from the report, the default answer to the question "Was random assignment to groups used?" is "No." The one table lacking this column concerns the completeness of statistical reporting; this will not need an "If Unreported" column.

Finally, due to the large number of possible combinations that result in a "Maybe No" answer, in some cases (like above) we had to omit that column of answers. A note is provided at the end of the table notifying the reader that this has occurred.

A glossary of terms is presented at the end of this report. The glossary is not complete at this time but does capture many of the most important terms. Also, some of the included terms require further operational specification, such as differential attrition. Finally, some terms require a definition specific to an area of research and will be provided by the Principal Investigator associated with each evidence report. For example, terms that need topic-area related definitions include: “important population and setting characteristics”, and “commonly-shared or theoretically-derived notions (about the content of interventions or approaches).”

Going From the Specific “Yes/No” Questions to the Eight Composite Scales

Below we provide the specific “Yes/No” questions and the eight composite questions. We also provide the algorithms applied to the answers to the specific questions that lead to answers to the eight composite level questions.

Composite Question #1: "Construct Validity - Intervention or Approach"

Was the *intervention or approach** properly defined?

- Yes, the *intervention or approach** was adequately described and it fully reflected commonly-held or theoretically derived ideas about what the intervention or approach should be (as defined by the PI of the Evidence Report).
- Maybe yes, at a minimum the *intervention or approach** was adequately described, and it at least largely reflected commonly-held or theoretically derived ideas about what the intervention or approach should be.
- Maybe no, the *intervention or approach** was described only as member of broader classes (across which significant variation in content can be expected).
- No, it is unclear what the *intervention or approach** was, or the intervention or approach did not reflect commonly-held or theoretically derived ideas about what it should be.

Minimum for inclusion in a WWC Evidence Report: "Maybe No"

Construct Validity – Intervention or approach	**Yes**	**Maybe Yes**	**Maybe Yes**	**Maybe Yes**	**No**	**No**	**If unreported:**
1.1a Was the *intervention or approach** described at a level which would allow its implementation by other implementers?	Yes	Yes	No	No	No	Yes or No	No
1.1b If no, was the *intervention or approach** a member of a broader class (across which significant variation in content can be expected)?	n/a	n/a	Yes	Yes	No	Yes or No	No
1.2a Does the *intervention or approach** fully *reflect commonly-held or theoretically derived ideas about what it should be** (as defined by the PI of the Evidence Report)?	Yes	No	No	Yes	Yes or No	No	No
1.2b If no, does the *intervention or approach** largely *reflect commonly-held or theoretically derived ideas about what it should be**?	n/a	Yes	Yes	n/a	Yes or No	No	
1.2c If no, does the *intervention or approach** somewhat *reflect commonly-held or theoretically derived ideas about what it should be**?	n/a	n/a	n/a	n/a	Yes or No	No	
1.3 Did the comparison group receive an alternative to the intervention that was meant to act as a placebo?	Yes	Yes or No	Yes or No	Yes or No	Yes or No	Yes or No	
1.4 Was the level of implementation measured or described?	Yes	Yes or No	Yes or No	Yes or No	Yes or No	Yes or No	

Note: Any pattern of answers not explicitly described results in a "Maybe No" answer to Composite Question #1.

Note: Terms in italics and followed by an asterisk are defined by the Principal Investigator of the Evidence Report.

Composite Question #2: "Construct Validity – Outcome Measures"

Was the outcome measure adequately defined?

- Yes, the report presented evidence that the outcome measure was properly defined.
- Maybe yes, the report did not present evidence that the outcome measure was properly defined, but it did appear to represent the content of interest.
- Maybe no, the outcome was described only as members of broader class (across which significant variation in content can be expected).
- No, it is unclear what the outcome was.

Minimum for inclusion in a WWC Evidence Report: "Maybe No"

Construct Validity -- Outcome Measure	Yes	Maybe Yes	Maybe No	No	If unreported:
2.1 Was there evidence of the *construct validity* of the outcome measure available either in the report or through other easily accessible documents?	Yes	No	No	No	No
2.2 Do items on the outcome measure appear to represent the content of interest to this Evidence Report (i.e., have face validity)?	Yes	Yes	No	No	No
2.3 If there was no evidence of construct validity and the report gave no indication of face validity, was the outcome measure a member of a broad class of measures (across which significant variation in content is to be expected)?	n/a	n/a	Yes	No	No

Note: Terms in italics and followed by an asterisk are defined by the Principal Investigator of the Evidence Report.

Composite Question #3: "Internal Validity – Selection"

Were the participants (e.g., students, schools) in the group receiving the intervention or approach comparable to the participants in the comparison group?

- Yes, participants were randomly assigned to conditions and there was no differential attrition or severe overall attrition.
- Maybe yes, EITHER random assignment was used but there was differential and/or several overall attrition OR although random assignment was not used, there was no attrition problem and reasonable steps were taken to make the groups comparable.
- Maybe no, random assignment was not used and although steps were taken to make the groups comparable, they did not appear to be adequate.
- No, it is unlikely that the participants in the groups were comparable.

Minimum for inclusion in a WWC Evidence Report: "Maybe Yes"

Internal Validity – Selection	**Yes**	**Maybe yes**	**Maybe yes**	**Maybe yes**	**No**	**If unreported:**
3.1 Was random assignment used to assign participants to intervention and comparison groups?	Yes	Yes	No	No	No	No
3.2 Was there *differential attrition** between intervention and comparison groups or *severe attrition* overall*?	No	Yes	No	Yes	Yes or No	No
3.3 If the groups were not demonstrably equivalent, were post-program procedures used that equated groups on a pretest of the outcome?	Yes or No	Yes or No	Yes or No	Yes	No	No
3.4 If the groups were not demonstrably equivalent, were pre-program procedures used that equated groups on a pretest of the outcome?	Yes or No	Yes or No	Yes	Yes or No	No	No
3.5 If the groups were not demonstrably equivalent, were post-program procedures used that equated groups on other important variables?	Yes or No	Yes or No	Yes or No	Yes	No	No
3.6 If the groups were not demonstrably equivalent, were pre-program procedures used that equated groups on other important variables?	Yes or No	Yes or No	Yes	Yes or No	No	No

Note: A pattern of answers to these questions that is not specifically identified results in a "Maybe No."
Note: Terms in italics and followed by an asterisk are defined by the Principal Investigator of the Evidence Report.

Composite Question #4: "Internal Validity – Contamination"

Was the study free of events that happened concurrently with the intervention or approach that confused its effect?

- Yes, concurrent processes and events that might be alternative explanations to a policy or practice effect have been ruled out.
- Maybe yes, there were no identified processes or events that could be alternative explanations, but some alternative explanations cannot be explicitly ruled out.
- [There is no "maybe no" answer for this question.]
- No, identifiable processes happening at the same time as the intervention or approach may have caused the effect.

Minimum for inclusion in a WWC Evidence Report: "Maybe Yes"

Internal Validity -- Intervention uncontaminated by other events	**Yes**	**Yes**	**Maybe Yes**	**No**	**No**	**No**	**If unreported:**
4.1 Was there evidence of a local history event?	No	No	No	Yes	No	Yes	No
4.2a Were the intervention and comparison groups drawn from the same *local pool**?	No	Yes	Yes	Yes or No	Yes or No	Yes or No	Yes
4.2b If yes, were intervention conditions known to study participants, providers, data collectors, and/or other authorities (e.g., parents, teachers, case managers)?	n/a	No	Yes	Yes or No	Yes or No	Yes or No	Yes
4.3 Did the description of the study give any other indication of the plausibility of other treatment contaminants?	No	No	No	No	Yes	Yes	No

Note: Terms in italics and followed by an asterisk are defined by the Principal Investigator of the Evidence Report.

Composite Question #5: "External Validity – Sampling"

Were *targeted participants, settings, outcomes, and occasions** included in the study?

- Yes, the *targeted participants, settings, outcomes, and occasions** are represented in the sample.
- Maybe yes, most *important characteristics of the participants, settings, outcomes, and occasions** are represented in the sample.
- Maybe no, although some important *characteristics of the participants, settings, outcomes, and occasions** are represented in the sample, many important targets are not.
- No, either the most *important characteristics of the target participants and settings** were not included, or the sampled participants were not part of the target population.

Minimum for inclusion in a WWC Evidence Report: "Maybe No"

External Validity – Sampling	**Yes**	**Maybe Yes**	**Maybe Yes**	**No**	**No**	**No**	**No**	**If unreported**
5.1a Did the sample contain participants with *the necessary characteristics to be considered part of the target population (for purposes of this Evidence Report)**?	Yes	Yes	Yes	Yes or No	Yes or No	No	Yes or No	No
5.1b If yes, was the sample randomly selected from the *target population**?	Yes	Yes	No	Yes or No	Yes or No	No	Yes or No	
5.1c If no, did the sample contain participants with a variety of *important characteristics* that could vary within the target population*?	n/a	n/a	Yes	Yes or No	Yes or No	No	Yes or No	Yes
5.2a Did the sampled settings contain all *important characteristics of the target setting* (for purposes of this Evidence Report)*?	Yes	No	No	Yes or No	No	Yes or No	Yes or No	No
5.2b If no, did the sampled settings contain some, but not all, *important characteristics of the target settings**?	n/a	Yes	Yes	Yes or No	No	Yes or No	Yes or No	No
5.3 Did the study measure the outcome at a *time appropriate for capturing the intervention's effect**?	Yes	Yes	Yes	No	Yes or No	Yes or No	Yes or No	Yes
5.4 Was the study conducted during the *time frame appropriate for the Evidence Report**?	Yes	Yes	Yes	Yes or No	Yes or No	Yes or No	No	No

Note: A pattern of answers to these questions that is not specifically identified results in a "Maybe No."
Note: Terms in italics and followed by an asterisk are defined by the Principal Investigator of the Evidence Report.

Composite Question #6: "External Validity – Testing within Subgroups"

Was the intervention or approach tested for its effectiveness within *important subgroups of participants, settings, outcomes, and occasions**?

- Yes, the *intervention or approach** was tested for its effectiveness on *targeted participants, settings, outcomes, and occasions**.
- Maybe yes, the *intervention or approach** was tested for its effectiveness within most *important subgroups of the participants and settings**.
- Maybe no, although the *intervention or approach** was tested for its effectiveness within some *important subgroups of the participants and settings**, many were left out.
- No, the *intervention or approach** was not tested for its effectiveness within most *important subgroups of the participants, settings, outcomes, and occasions**.

Minimum for inclusion in a WWC Evidence Report: Not applicable -- any score is acceptable

External Validity – Testing within subgroups	**Yes**	**Maybe Yes**	**Maybe Yes**	**Maybe Yes**	**Maybe Yes**	**No**	**If unreported:**
6.1a Was the intervention or approach tested for effectiveness within *all important subgroups of participants* (for purposes of this Evidence Report)*?	Yes	No	Yes	Yes	Yes	No	No
6.1b If no, was the intervention or approach tested for effectiveness within some, but not all, *important subgroups of participants**?	n/a	Yes	n/a	n/a	n/a	No	No
6.1c If no, was the intervention or approach tested for effectiveness within any *important subgroups of participants**?	n/a	n/a	n/a	n/a	n/a	No	No
6.2a Was the intervention or approach tested for effectiveness within all *important subgroups of settings* (for purposes of this Evidence Report)*?	Yes	Yes	No	Yes	Yes	No	No
6.2b If no, was the intervention or approach tested for effectiveness within some, but not all, *important subgroups of settings**?	n/a	n/a	Yes	n/a	n/a	No	No
6.2c If no, was the intervention or approach tested for effectiveness within any *important subgroups of settings**?	n/a	n/a	n/a	n/a	n/a	No	No
6.3 Was the intervention or approach tested for its effectiveness across *classes of outcomes**?	Yes	Yes	Yes	No	Yes	No	No
6.4 Was time of measurement (relative to the end of the intervention) tested as an influence on the treatment effect?	Yes	Yes	Yes	Yes	No	No	No

Note: A pattern of answers to these questions that is not specifically identified results in a "Maybe No."
Note: Terms in italics and followed by an asterisk are defined by the Principal Investigator of the Evidence Report.

Composite Question #7: "Statistical Validity – Effect Size Estimation"

Were the effect sizes accurately estimated?

- Yes, the effect sizes appear to be accurately estimated.
- Maybe yes, there was some evidence of statistical issues that may have caused the effect sizes to be inaccurately estimated, but the likely impact on inferences was minimal.
- Maybe no, there was evidence of statistical issues that may have caused the effect sizes to be inaccurately estimated.
- No, the assumption of statistical independence was not met, and dependence was not accounted for in the effect sizes.

Minimum for inclusion in a WWC Evidence Report: "Maybe No"

Statistical Conclusion Validity – Effect Size	**Yes**	**Maybe Yes**	**Maybe yes**	**Maybe No**	**No**	**If unreported:**
7.1 Was the assumption of independence met, or could dependence (including dependence arising from clustering) be accounted for in estimates of effect size and their standard errors?	Yes	Yes	Yes	Yes	No	No
7.2 Did the *statistical properties of the data** (e.g., distributional and variance assumptions, if any, presence of outliers) allow for valid estimates of the effect sizes?	Yes	No	Yes	No	Yes or No	Yes
7.3 Was the *sample size adequate** to provide a sufficiently precise estimate of the effect size?	Yes	Yes	No	No	Yes or No	No
7.4 Were the outcome measures sufficiently reliable to allow a sufficiently precise estimate of the effect size?	Yes	Yes	Yes	Yes or No	Yes or No	Yes

Note: Terms in italics and followed by an asterisk are defined by the Principal Investigator of the Evidence Report.

Composite Question #8: "Statistical Validity – Completeness of Reporting"

Were the statistical tests adequately reported?

- Yes, the statistical tests were completely reported.
- Maybe yes, sufficient statistical information was reported to allow imprecise effect sizes to be calculated for most measured outcomes.
- Maybe no, effect sizes could not be calculated for most outcome measures.
- No, sample size was not reported, OR the direction of the effect could not be discerned for most outcome measures.

Minimum for inclusion in a WWC Evidence Report: "Maybe No"

Statistical Conclusion Validity -- Completeness of Reporting	**Yes**	**Maybe Yes**	**Maybe No**	**No**	**No**
8.1 Was sample size reported or could it be derived or estimated from statistical information presented?	Yes	Yes	Yes	Yes or No	No
8.2 Could directions of effect be identified for most *important measured outcomes**?	Yes	Yes	Yes	No	Yes or No
8.3a Could effect sizes be derived (e.g., from variances/covariances, translated from other metrics) for most measured outcomes?	Yes	No	No	Yes or No	Yes or No
8.3b If no, could effect sizes be roughly estimated from imprecise significance levels?	n/a	Yes	No	Yes or No	Yes or No

Note: Terms in italics and followed by an asterisk are defined by the Principal Investigator of the Evidence Report.

Going From the Eight Composite Scales to the Four Global Scales

Below we provide the answers to the four global questions. We also provide the algorithms applied to the answers to the eight composite questions that lead to each global level answer.

Global Question #1: "Construct Validity"

Were the intervention or approach and outcome measures properly defined?

- Yes, the intervention or approach and the outcome measures were properly defined
- Maybe yes, at a minimum the intervention or approach at least largely reflected commonly-held or theoretically-derived ideas about what it should be, and the outcome measures appeared to measure the content of interest.
- Maybe no, the intervention or approach and/or the outcome measures were described only as a members of broader classes (across which significant variation in content is to be expected).
- No, it is unclear what was done in the study.

	Yes	**Maybe Yes**	**Maybe Yes**	**Maybe Yes**	**No**	**No**
Was the intervention or approach properly defined (as defined by the PI of the Evidence Report)?	Yes	Yes	Maybe Yes	Maybe Yes	No	Yes or No
Was the outcome properly defined (as defined by the PI of the Evidence Report)?	Yes	Maybe Yes	Yes	Maybe Yes	Yes or No	No

Note: A pattern of answers to these questions that is not specifically identified results in a "Maybe No."

Global Question #2: "Internal Validity"

"Was the intervention or approach the cause of the change in the outcome?"

- Yes, the major alternative explanations (selection and contamination) have been ruled out.
- Maybe yes, although steps were taken to make the groups comparable, it is possible that attrition or a lack of randomization caused them to differ somewhat.
- Maybe no, random assignment was not used to make groups comparable, and it seems likely that any steps taken to make them comparable were inadequate.
- No, it is unclear what might have caused the difference.

	Yes	**Maybe Yes**	**Maybe Yes**	**Maybe Yes**	**No**
Were the participants in the group receiving the intervention or approach comparable to the participants in the comparison group?	Yes	Maybe Yes	Yes	Maybe Yes	No
Was the study free of events that happened concurrently with the intervention or approach that confused its effect?	Yes	Yes	Maybe Yes	Maybe Yes	No

Note: A pattern of answers to these questions that is not specifically identified results in a "Maybe No."

Global Question #3: "External Validity"

Was the intervention or approach tested on relevant participants (for example, students, schools) and environments (for example, classrooms, occasions)?

- Yes, the proper targets were included and the effect of the intervention or approach was tested within targets.
- Maybe yes, at least some important targets were included in the study, and the intervention or approach was tested within some these targets.
- Maybe no, many important targets were not included in the study, and/or the intervention or approach was rarely tested within targets.
- No, the accessed sample was not part of the target population.

	Yes	**Maybe Yes**	**Maybe Yes**	**Maybe Yes**	**No**
Were *targeted participants, settings, outcomes, and occasions** included in the study?	Yes	Maybe Yes	Yes	Maybe Yes	No
Was the intervention or approach tested for its effectiveness within *important subgroups of participants, settings, outcomes, and occasions**?	Yes	Yes	Maybe Yes	Maybe Yes	Yes or No

Note: A pattern of answers to these questions that is not specifically identified results in a "Maybe No."

Global Question #4: "Statistical Conclusion Validity"

Could accurate effect sizes be derived from the study report?

- Yes, the statistical results were adequately reported and the effect sizes accurately estimated.
- Maybe yes, either the statistical results were only imprecisely reported, or there was evidence that the effect sizes may have been inaccurately estimated.
- Maybe no, there were important problems with the reporting of the statistical results and/or the estimation of the effect size.
- No, statistical results were not reported for most measured outcomes, and/or there was evidence that the effect sizes may have been inaccurately estimated.

	Yes	**Maybe Yes**	**Maybe Yes**	**No**	**No**
Were the effect sizes accurately estimated?	Yes	Maybe Yes	Yes	Yes or No	No
Were the statistical tests adequately reported?	Yes	Yes	Maybe Yes	No	Yes or No

Note: A pattern of answers to these questions that is not specifically identified results in a "Maybe No."

GLOSSARY OF TERMS

1:1 correspondence: Some measures are of interest in their own right, and not necessarily as measures larger constructs. For example, a program intended to increase graduation rates among at-risk high school students will likely measure whether program participants actually graduated from high school. Variables such as these do not need to demonstrate construct validity, even though they may not be perfectly measured or consistently defined.

Appropriate time for capturing the intervention's effect: Occasionally, programs may be evaluated at an inappropriate time. For example, the program might be evaluated before the intervention might reasonably be expected to have an effect on participants.

Assumption of Independence: The assumption that the errors of an experiment are independent of one another (Rosenthal & Rosnow, 1991, p. 315)

Attrition: Loss of participants that occurs after group assignment has taken place (also called *mortality*) (Shadish, Cook, & Campbell, 2002, 2002, p. 505).

Comparison Group: In an experiment, a group that is compared with a treatment group and that may receive either an alternate intervention or no intervention (Shadish, Cook, & Campbell p. 506)

Compensatory process: A process by which the contrast between the control/comparison group and the treatment group is lessened because (a) the treatment is administered to the control group (compensatory equalization), (b) the treatment is removed from the treatment group (compensatory deprivation), (c) the control group performs better to match achievements of the treatment group (compensatory rivalry), or (d) the control group performs poorly because it feels demoralized (resentful demoralization) (Shadish, Cook, & Campbell, 2002, p. 79-80). For example, (a) a teacher in the control groups adopts the treatment before the study is over, (b) a teacher drops the treatment because it is not working prior to the end of the study, (c) a control teacher increases the pace of instruction to match the effect of the treatment approach in another class, and (d) a control teacher feels neglected and teaches with less enthusiasm.

Construct validity: The degree to which inferences are warranted from observed characteristics of participants, the intervention or approach, and outcomes sampled within a study to the constructs these samples represent. For example, an outcome measure may demonstrate construct validity by establishing convergent and/or discriminant validity, by behaving in theoretically predicted ways in experimental settings, or having a 1:1 correspondence between the construct and the measurement, among others.

Convergent validity: The idea that two measures of the same thing should correlate with each other (Shadish, Cook, & Campbell, 2002, p. 506)

Demonstrably equivalent: When a report provides evidence that intervention and comparison groups happened to be well matched on a pretest of the outcome and on the other important variables (as defined by the PI of the Evidence Report) after attrition (if any).

Direction of Effect: The direction (positive or negative) that the outcome measure has (relative to the comparison group) on the intervention group's standing on the outcome variable. An imprecise

estimate of the effect of the intervention or approach can be estimated from a set of studies that provide (a) the direction of the effect and (b) sample size, providing that not all directions are of the same valence.

Discriminant validity: The notion that a measure of A can be discriminated from a measure of B, when B is thought to be different from A; discriminant validity correlations should be lower than convergent validity correlations (Shadish, Cook, and Campbell, 2002, p. 507)

Effect size: The strength (or magnitude) of the relationship in the population, or the degree of departure from the null hypothesis. (Rosenthal & Rosnow, 1991, p. 42)

Equate: Procedures used to make groups more comparable.

Face validity: The extent to which an instrument "looks like" it measures what it is intended to measure (Nunnally, 1967, p. 99)

Properly implemented: Every participant in each condition fully receives the intervention, and fully complies with the intervention to which they were assigned and receives no other intervention. The intervention is not tailored to particular participants (Shadish, Cook, & Campbell, 2002, p. 315)

Implementation description: The activities, both intended and unintended, that did and did not occur as part of the treatment conditions; includes treatment delivery, treatment receipt and treatment adherence (Shadish, Cook, & Campbell, 2002, p. 508, 316)

Local history event: An event occurring between the beginning of the treatment and the posttest within the context of the treatment, outcome, time, setting, and persons studied that could have produced the observed outcome in the absence of the treatment (Shadish, Cook, & Campbell, 2002, p. 508, 316)

Local pool: Persons having a geographical or other contextual relationship with the sampled participants. For example, students in the same classroom or school might be considered part of the same local pool.

Multiple occasions: Outcomes are measured at multiple points in time, relative to the end of the intervention.

Placebo: An intervention that does not include the presumed active ingredients of the treatment. For example, a placebo in education would be used to help rule out administrator expectancy effects and novelty and disruption effects

Intervention or approach: The focus of a WWC Evidence Review.

Positive evidence: In this context, positive evidence refers to an unambiguous identification of a problem that occurred during the implementation of the intervention or approach.

Post-intervention procedures used to equate groups: In this context, post-intervention refers to equating procedures that were applied after the intervention ended. For example, these procedures would include matching and statistical adjustment. By definition, post-intervention procedures equate

groups on measured characteristics after attrition has already occurred. It does not control for the effects of unmeasured characteristics that may have led to attrition.

Pre-intervention procedures used to equate groups: In this context, pre-intervention refers to equating procedures that were applied before the intervention started. For example, these procedures would include matching and statistical adjustment. Pre-intervention procedures used to equate groups do not control for any possible effects differential attrition.

Random assignment: In an experiment, any procedure for assigning participants to conditions based on chance, with every participant having a nonzero probability of being assigned to each condition (Shadish, Cook, & Campbell, 2002, p. 511)

Random selection: More general term that is sometimes used synonymously with either random sampling or random assignment in different contexts (Shadish, Cook, & Campbell, 2002, p. 511)

Sample: The subset of the population for whom we have obtained observations (Rosenthal & Rosnow, p. 628)

Sample size: The number of observations upon which the effect size is based.

Self-selection: When participants decide the condition they will be in (Shadish, Cook, & Campbell p. 512). For example program volunteers are compared to non-volunteers or people completing the treatment are compared to people not completing the treatment.

Significance levels: The level of alpha; also called *p value* (Rosenthal & Rosnow, p. 629). Imprecise (and conservative) effect sizes can be calculated when all that is known is the sample size and that a given effect was significant at $p < .05$.

Target population: The complete collection of individuals for whom the intervention or approach is designed for. This is defined by the review team separately for each topic.

Other treatment contaminants: Events or processes other than those listed individually that might be confused with a treatment effect. These would include interactions between contamination threats and group membership, for example, when multiple treatments occur and the treatment effect only appears when other treatments are present.

Unit of analysis: The unit at which statistical analyses were performed. This may differ from the unit of assignment or observation.

Unit of assignment: The level at which participants (e.g., students, schools) were assigned to conditions.